What is a Language? What is a Dialect?*

Hizniye Isabella Boga
Eberhard Karls Universität Tübingen
isabella.boga@uni-tuebingen.de

The current study shows how to distinguish dialects from languages. This distinction was found with the help of the Needleman-Wunsch algorithm with a weighted scorer system of PMI distances and the Levenshtein distance. The study focused on the Romance language family, especially languages of Italy. The means used in order to identify groupings in the data were mixture models and the *k*-means clustering.
The results support the hypothesis of bearing two thresholds which divide language-language pairs, language-dialect pairs and dialect-dialect pairs into three distinct clusters. These clusters were found with the Needleman-Wunsch algorithm with normalised and divided (NWND) scores and an additional scorer system of PMI distances. Furthermore, I also used Levenshtein Distances Normalised and Divided (LDND) for comparative reasons.
The suggested thresholds differentiated between the two methods. The threshold by the NWND method are 4.49 for distinguishing dialect-dialect pairs from language-dialect pairs and a threshold of 2.54 in order to distinguish dialect-language pairs from language-language pairs. For the LDND method the cut off-points are 0.37 to distinguish dialect-dialect pairs from dialect-language pairs, 0.58 to distinguish close dialect-language varieties from distant dialect-language varieties and 0.7 to distinguish distant dialect-language varieties from language-language pairs.

1. Introduction

The question of what a dialect is in opposition, relation or contrast to a language has been answered many times in many different manners. One common way of definition is that a dialect, to put it in broad terms, is a subclass of a language. This means that Palatine is a dialect of the German language, Mancunian is a dialect of English and Kurmanji is a dialect of Kurdish.

* This research was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 834050).

A language hence is a "collection of mutually intelligible dialects" according to Chambers et al. (1998). This definition would go into the right direction if there was not the problematic case of politics, geography or religion. Languages often get designated due to one of the aforementioned reasons Chambers et al. (1998). This can be seen in the case of Urdu and Hindi. They are in fact the same language with a different writing system and a different distribution of loanwords. Speakers of the two languages understand each other, which means that Hindi and Urdu are mutually intelligible King (2001). According to the general definition, this would make them two dialects of the same language, but due to political reasons and because they are allocated to two different countries, they are considered two distinct languages. The same case holds for Norwegian and Swedish. Speakers of both languages can understand each other fairly well but because they are associated with different countries, they also are considered two distinct languages.

Another case is Chinese which is often considered a single language whereas it actually consists of at least 7 different languages with major dialectal variation. And here again, due to political reasons and also due to a common writing system, Chinese is considered one language.

This is a major point I want to put emphasis on, namely that one should not forget that languages and dialects are mostly defined as what they are due to political reasons and not due to linguistic properties. Chambers et al. (1998) pinpoints that unless we do not want to change our assumption of what a language is, we have to radically assume that a language is not a particularly linguistic notion at all. The foremost important point for the purpose of this thesis is to broaden the already fixed definitions of what a dialect is and what a language is. The idea is to approach this field from a different perspective and examine the languages, or, to be precise, the varieties of a language from a new angle. What is taken into account are varieties of languages and associated dialects.

Another important aspect that is often disregarded is mono-directional intelligibility when taking intelligibility as a measuring factor for classification. This can be observed with the case of Danish and Norwegian. It is said that Danes understand Norwegians better than the other way round Gooskens (2006). There are many other aspects when it comes to intelligibility, for instance the educational degree of the speakers, social and socio-linguistic aspects, geographical and historical circumstances etc., but these are disre-

garded for my purposes as that would go beyond the scope of this article.

Some definitions would need to be revised if the classification of "languages" was purely based on linguistic features. One typical example would be German and its varieties on the one hand and Dutch on the other hand. German and Dutch are considered to be two fully-fledged languages. The German dialect Bavarian is considered a variety of German. Nevertheless, it is much more difficult (and sometimes even impossible) for someone from the German side of the Dutch border to understand any German variety like Bavarian (which is considered to be closer, as it is a dialect of the same Standard variety) than Dutch even though this is considered to be a language by itself. Hence my claim is that one has to consider varieties in general on a gradual scale rather than as categorical entities. This means that a varieties' affiliation to one or another "language" is not clear-cut but rather ambiguous due to them not being discrete entities.

Amongst dialectological approaches, various scholars have strived to find objective measurements for different cases in dialect research. This ambition is the seed of the emergence of dialectometry. Jean Séguy, a pioneer in the field of dialectometry, pursued the goal of finding an objective measure in order to find differences in dialects. He is also, amongst others, considered the father of dialectometry Wieling & Nerbonne (2015).

In 1995, Brett Kessler laid the foundation of what was to become a successful application of bio-informatic methods in linguistics. He was the first one to apply the Levenshtein distance to wordlists from Irish Gaelic to measure dialectal differences and infer dialect groupings Kessler (1995).

The first book in dialectometry was published by Hans Goebl in 1982 Goebl (1982). His work revolved around Romance languages and also introduced statistics and cluster analysis into the field of dialectology. This approach can be seen as the beginning of the shift from dialectology to dialectometry and hence the establishment of the field.

From there on, various researchers in the field of dialectology used objective measurements for cases like dialect groupings or measuring differences in pronunciations with string similarity measurements Heeringa (2004). Amongst the aforementioned publications, there are many others such as those of John Nerbonne, Peter Trudgill, J. K. Chambers and many more, that have contributed immensely to the field of either dialectology and/or dialectometry.

The idea of this project is to find objective measurements to differentiate groups of dialects from groups of languages in the Romance language family. As mentioned before, the idea of measuring this distance has been done exhaustively with the *edit distance* (see section 3 for further Information), as for instance in Wichmann (2019). That is why, for the sake of comparison, I use the Levenshtein distance as well. However, my main focus will be on the Needleman-Wunsch method (for further information on both methods please consult section 3). Both these algorithms measure distances between strings and will be explained further in section 3.

For this purpose, I measure the distance of words between language pairs and will infer thresholds for objective boundaries. These measurements are performed with the already mentioned *edit distance*, or Levenshtein Distance, and with the Needleman-Wunsch algorithm which is a further aspect building up on Wichmann (2019). One adjustment I performed on the method was to include weighted alignments in form of a scorer system which makes out the biggest and most significant difference to previous studies in the field. In order to analyse the data accordingly, I apply clustering methods.

With this work I want to elaborate the objective threshold between pairs of languages and pairs of dialects by means of distance measurements and mixture models. My hypothesis is to find an objective measurement between language-language pairs, language-dialect pairs and dialect-dialect pairs rather than finding a bipartite system as proposed in Wichmann (2019). Furthermore, I expect the Needleman-Wunsch algorithm with the weighted scorer system to work more precisely than the Levenshtein distance for the task of finding the distinction between languages and dialects. This hypothesis is based on the assumption that the scorer system in the Needleman-Wunsch method will give room for a more fine-grained distance analysis.

1.1. A Brief Sketch of Italian Dialects

Despite having analysed a large number of Romance languages in terms of the methods I used, the focus point in the forthcoming sketch will be the Italian varieties.

Standard Italian, the official language of Italy, is spoken by about 60 million people inside Italy and by another 1.3 million people in other European countries. Outside Europe, it is spoken by roughly 6 million people in North

and South America Clivio et al. (2011).

Standard Italian is a direct descendant of the *volgare* (the language of the common folk), or Vulgar Latin, and became a separate language around 1000 C.E. This form of Latin is the spoken variety used by the common people, hence the term "Vulgar Latin" Clivio et al. (2011). In the early 1200s, the Sicilian *volgare* was the common variety used in literature due to Sicily functioning as the centre of European cultural life Clivio et al. (2011). It was not until 1250 that this centre of cultural fortress shifted to Tuscany and hence the Tuscan *volgare* became the basis for the common language Clivio et al. (2011) due famous writers like Boccaccio, Alighieri etc. and trade.

Jumping 600 years forward, two major isoglosses divide the country — the *La Spezia-Rimini* line and the *Rome-Ancona* line which, according to Maiden & Parry (1997), dividing the country into three dialectal regions:

Name of Region		
Northern Dialects	Central Dialects	Southern Dialects
1. The Gallo-Italic dialects: Piedmontese, Lombard, Ligurian, Emilian and Bolognese 2. The Venetian dialect, which is spoken in the Italian Tyrol and parts of Dalmatia and Istria in addition to the Venetian area itself	1. Tuscan 2. Corsican 3. North Sardinian 4. Romanian, Umbrian and Merchesan	1. The "upper" dialects: Campanian, Abruzzese, Molisano, North Apulian, Neapolitan, Lucanian 2. The "extremity" dialects: Sicilian, Calabrian, South Apulian and Salentino

Table 1. The Three Distinct Areas According to Clivio et al. (2011) and Maiden & Parry (1997)

Sometimes a fourth group is added called *Transitional Dialects* and can be found on the Adriatic coast between Romagna and the Marche region, but the tripartite classification seems to prove more popular.

Consider the maps in 1 and 2. They depict the relation of one variety towards every other variety in the data set by colour. These relations are the results of the distance measurements between every variety covered in the Romance data set. The darker the colour, the lower the distance, hence the closer the varieties.

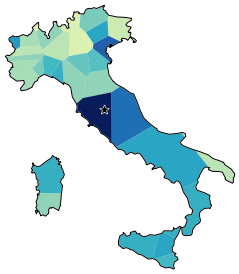


Figure 1. Distances between Standard Italian and the Rest of Italy

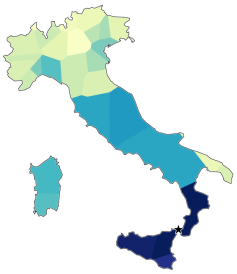


Figure 2. Distances between Messinese Sicilian and the Rest of Italy

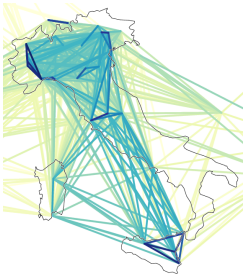


Figure 3. Distances between Every Datapoint in the Republic of Italy

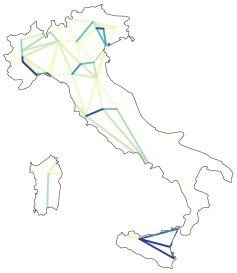


Figure 4. Strongest Distances between Every Datapoint in the Republic of Italy

We can see that there is a gradual digression from Messinese Sicilian (2) to every other variety in Italy. Almost the same holds for Standard Italian (1). Nonetheless, as the influence of Standard Italian is by far the largest in Italy, we can see a somewhat more meddled up picture. So it holds that Standard Italian is still closer to some dialects in the north of Italy, whereas Messinese Sicilian shows a gradual change from its closest neighbours to distant varieties more clearly. This is a natural phenomenon called dialect

continuum, which is well portrayed in 2. A dialect continuum can be seen as a mutual intelligibility on a chain, where each chain link is one variety. Intelligibility means a at least partial understanding of one another. Consider four different varieties A, B, C and D that are spoken in four different but adjacent locations. While variety A and B are mutually intelligible and variety B and C also are mutually intelligible, one would already recognise divergences between variety A and C but they would probably be still mutually intelligible. If one goes further away from location A and compares that variety to variety D for instance, it would most likely be already very distant and hard to understand for speakers of variety A, whereas for speakers of variety C it would still be well understandable. This pattern can be observed in the figures 1 and 2. Again, the darker the shade of colour the closer is the variety. In 2 it is visible that the shade of colour brightens with the geographic distance except for two darker patches in the north of Italy and one bright patch in the south east. Nonetheless, the concept of a dialect continuum can be nicely shown with this data. As already mentioned, Standard Italian has a major influence on other varieties of Italy and the dialect continuum can not be as clearly seen as with Messinese Sicilian.

In the figures 3 and 4 we can see distances between every data point in the Republic of Italy. The darker the colour, the lower the distance. Close varieties bear little to no distance whereas distant varieties usually have a larger distance to each other. This is also to be expected to be seen in the results.

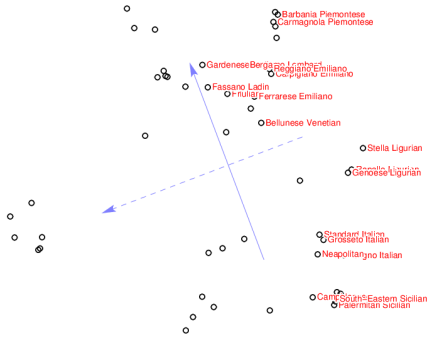


Figure 5. MDS Plot with Some Varieties of Italian

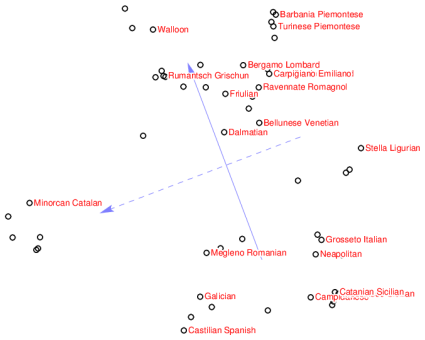


Figure 6. MDS Plot with Some Varieties of the Romance Languages

In plot 5 we see the distances between varieties on a Multidimensional Scaling plot. For the sake of readability not every variety from the Romance

data set is included by name on the MDS plot but only as a data point. The plot in 5 shows only Italian varieties, the plot in 6 also shows other varieties in order to have a reference point.

It is clearly visible that the dialects of Italy are spread along an axis. Catanian Sicilian is, for example, very far away from Barbania Piemontese or Bellunese Venetian. Compare this distance to that of Catanian Sicilian to Minorcan Catalan. They seem to be as far away as the two dialects from the same country. Geographically close varieties seem to be also closer to each other on the MDS plot. These distances and similarities will be evaluated in the upcoming section.

2. Data

The data used for this study is the Romance data set from the Global Lexicostatistical Database Starostin (2011). The data covers 58 Romance varieties with 110 concepts for each language. The concept list is based on the idea of a basic concept list according to Morris Swadesh and contains fundamental concepts that are considered very stable and unlikely to be borrowed Swadesh (1971).

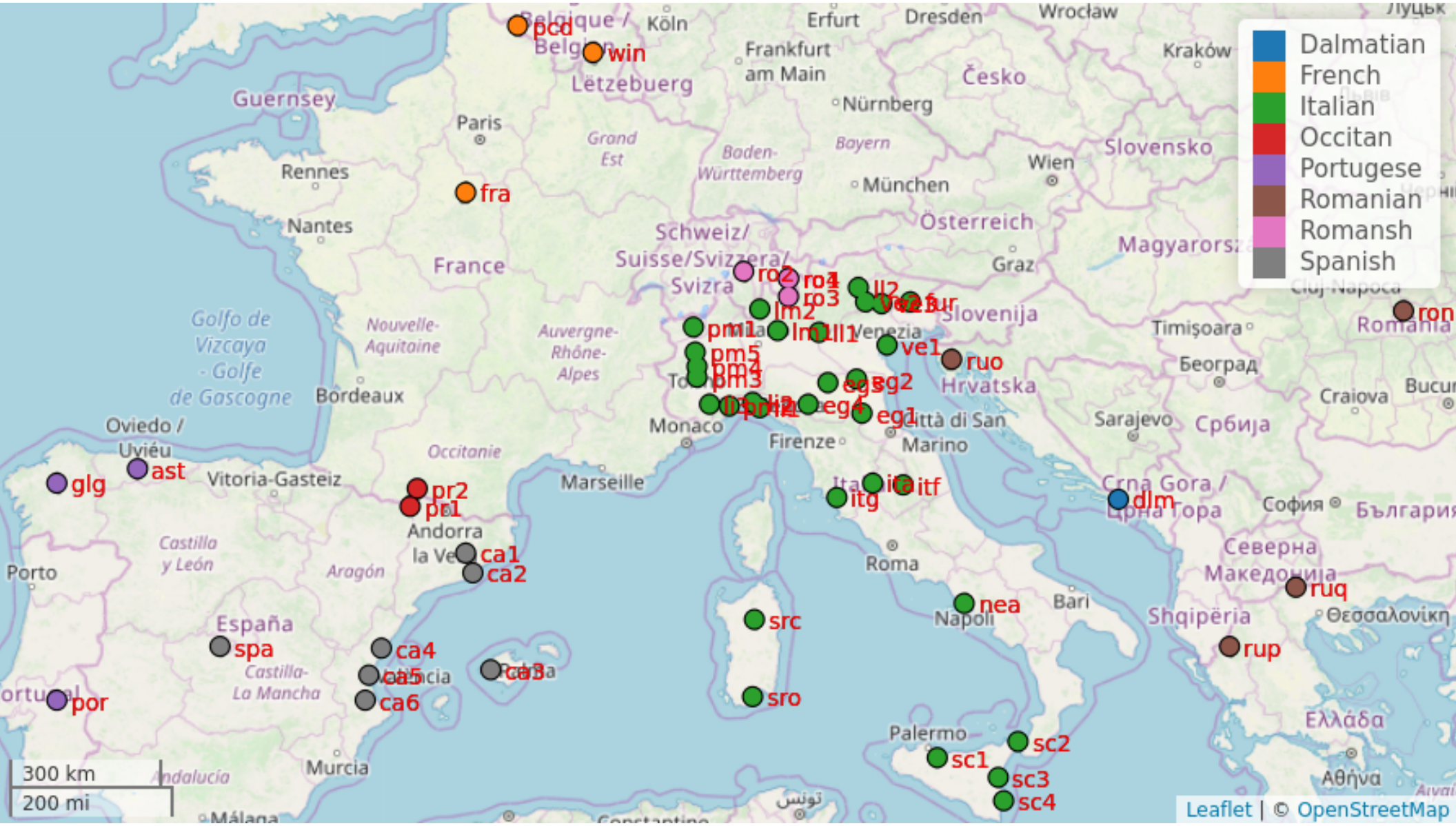
In order to make use of the data, some changes had to be conducted. The first adjustment was the conversion of IPA characters into ASJP characters. The reason for this change is the scorer I used for the Needleman-Wunsch algorithm which uses ASJP characters. This scorer system are the PMI distances kindly provided by Gerhard Jäger and were coded in ASJP characters. Hence the adaptation of the database had to be performed to be able to match the characters. The conversion was reproduced based from Brown et al. (2008).

Another adjustment that had to be implemented was the assignment of new, individual codes. Every standard language has an individual ISO-code but in this case, as I was dealing with dialects, some varieties do not have this individual assignment. Consequently, I had to alter the existing ISO-codes in order to identify every variety unambiguously. The corresponding new codes can be seen in the Appendix.

Originally, it included synonyms for some languages for only a few words but these were excluded as they were not regularly distributed and only covered a handful of words in addition to the given words in the regular data set. This means that for 42 languages optional synonyms were offered for only very

few concepts (a maximum of synonyms per concept for every language), but in most of the cases the entries were missing. Consequently, these rows (about 0.611 % of the entire data) were deleted from the data set and not considered in the analysis.

The map on the following page shows every language as a data point except for Archaic Latin, Classical Latin, Old Italian and Old French as these can not be allocated to a specific location. Nonetheless, they were included in the analysis. The colours in this map refer to a phylogenetic grouping of the varieties to their "Standard Variety" according to the leading literature.


Dalmatian
French
Italian
Occitan
Portugese
Romanian
Romansh
Spanish
pcd
win
fra
ro2
ro4
ro3
lm2
pm1
lm1
ll1
ll2
ve1
ruo
pm5
pm4
pm3
eg3
eg2
eg4
eg1
pm2
ll3
ll4
itg
itf
dlm
ron
pr2
pr1
ca1
ca2
ca4
ca5
ca6
ca3
spa
por
glg
ast
src
sro
nea
sc1
sc2
sc3
sc4
rup
ruq
Guernsey
Paris
Rennes
Nantes
France
Bordeaux
Golfo de Vizcaya - Golfe de Gascogne
Nouvelle-Aquitaine
Auvergne-Rhône-Alpes
Grand Est
Baden-Württemberg
Bayern
Schweiz/ Suisse/Svizzera/ Svizra
Monaco
Firenze
Marseille
Occitanie
Vitoria-Gasteiz
Oviedo / Uviéu
Porto
Castilla y León
Aragón
Castilla-La Mancha
España
Andorra
Murcia
Málaga
Roma
Napoli
Bari
Palermo
Venezia
Città di San Marino
Graz
Wien
München
Nürnberg
Frankfurt am Main
Köln
Erfurt
Dresden
Wrocław
Kraków
Česko
Slovensko
Österreich
Magyarország
Slovenija
Hrvatska
Srbija
Crna Gora / Црна Гора
Shqipëria
Ελλάδα
Sarajevo
Timişoara
Craiova
Beograd
София
Северна Македонија
Θεσσαλονίκη
Αθήνα
300 km
200 mi
Leaflet | © OpenStreetMap

3. Methods

The methods used in order to find a threshold between language-language, language-dialect and dialect-dialect pairs for the purpose of differentiating a dialect from a language are based on distance and similarity measurements. These measurements are realised with the Needleman-Wunsch method and the Levenshtein-Distance. Both methods are implemented in the *LingPy version 2.6.4* from November 26, 2018. *LingPy* offers "modules for sequence comparison, distance analyses, data operations and visualization methods in quantitative historical linguistics" List et al. (2017).

(1) The edit distance between two strings is defined as the minimum number of edit operations - insertions, deletions, and substitutions - needed to transform the first string into the second. For emphasis, note that matches are not counted. Gusfield (1997)

In order to calculate the distance D(i,j) between two sequences, one needs to opt for the minimal value in either of the three calculations given in the equation Levenshtein Distance Matrix Filler, Levenshtein (1966). The Levenshtein matrix D(i,j) is built up as suggested in the following:

1. Matrix construction D(i,j).
2. Initialisation of D(0,0) = 0.
3. Fill the matrix from the top left corner to the bottom right corner recursively.
4. Traceback of optimal alignment.

The matrix is filled depending on either dealing with a match, insertion, substitution or deletion. Consider the formula in Levenshtein Distance Matrix Filler, Levenshtein (1966):

(2)

(Levenshtein Distance Matrix Filler, Levenshtein (1966))

$$D(i,j) = min \begin{cases} D(i-1,j)+1 \\ D(i,j-1)+1 \\ D(i-1,j-1)+t(i,j) \end{cases}$$

A match and a mismatch would be computed with $D(i-1, j-1)+t(i, j)$, a deletion is computed with $D(i-1, j)+1$ and $D(i, j-1)+1$ corresponds to the calculation of an insertion Levenshtein (1966).

The Needleman-Wunsch algorithm, as already mentioned in the preceding section, is a global alignment and belongs to the Dynamic Programming algorithms by Saul B. Needleman and Christian D. Wunsch Needleman & Wunsch (1970).

The idea of dynamic programming is to "find an approach for the solution of complicated problems that essentially works the problem backwards" List (2013). An alignment is built up “using previous solutions for optimal alignments of smaller subsequences” List (2013), Durbin et al. (1998) rather than aligning two sequences in every possible manner and *then* picking the best score.

The Needleman-Wunsch algorithm finds the optimal alignment between two strings and is central to computational sequence analysis. The leading idea is to "build up an optimal alignment of smaller subsequences where the value F(i,j) is the score of the best alignment between the initial segments" Durbin et al. (1998). F(i,j), which is the score of the best alignment between the initial segments, is build up recursively:

1. Matrix construction F(i,j).
2. Initialisation of F(0,0) = 0.
3. Fill the matrix from the top left corner to the bottom right corner recursively.
4. Traceback of optimal alignment.

There are three possible ways to calculate the best score for F(i,j) for an alignment with x_i, y_i, where the best score up to (i,j) will be the largest of the three options given in the Needleman-Wunsch Matrix Filler:

(3)

(Needleman-Wunsch Matrix Filler)

$$F(i, j) = max \begin{cases} F(i-1, j-1) + s(x_i, y_i) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Durbin et al. (1998)

x_1 could be aligned to y_1; consequently, this equation would hold $F(i-1, j-1) + s(x_i, y_i)$. x_1 could be aligned to a gap in which the case of $F(i-1, j) - d$ would hold or y_1 could be aligned with a gap for which $F(i, j-1) - d$ would hold. The calculation will be done repeatedly to fill the matrix until the bottom right-hand corner is reached.

The final step is the trace back for the best alignment. The result of the computation will be traced back through the matrix. As the goal of the algorithm is to find the best alignment (which is represented by the highest possible value), it will always trace back along the line of the highest values in the matrix. This trace back can be done, in case of a matched alignment, diagonally and, in case of a gap, upwards and to the left Durbin et al. (1998). Following the trace back, one would reach the initial starting point F(0,0).

The calculated distances between a word pair with the same meaning is the absolute distance/similarity measurement. I will also refer to this as the diagonal score. In order to rule out negative similarity scores produced by the Needleman-Wunsch method, a constant was added to every score. This constant was determined by the lowest value which was -16.6. After this modification, every value was 0 or higher.

The absolute distance can be higher or lower according to the word length. To compensate for this, both Needleman-Wunsch scores and Levenshtein Distances were divided by the length of the longest word in a word-word comparison. Another factor are sound inventories. The chance to hit a high similarity score with Needleman-Wunsch or a low distance measurement with Levenshtein is also driven by the size of sound inventories. Jäger (2014) states that "if two languages have small and strongly overlapping sound inventories, the number of chance hits is high as compared to a language pair with large and dissimilar sound inventories." In order to eliminate this, the off-diagonal score was computed, which is the comparison of every word in one language with every word in another language without regarding their meaning. The difference to the diagonal score is the calculation is carried out regardless of the meaning. This entails a comparison of every single word of Language a with every single word of Language B. By definition, this is called "Levenshtein Distance Normalized and Divided" (LDND) and "Needleman-Wunsch Score Normalized and Divided" (NWND).

A subtlety that was added are the PMI scores functioning as a scorer. Point-

wise Mutual Information (PMI) compares the joint probability of observing *x* and *y* with the probabilities of observing them independently, where *x* and *y* are two points (or words). The formal definition is given in Pointwise Mutual Information. The joint probability *P(x,y)* will be much larger than chance *P(x)* and *P(y)* if there is a genuine association between these two points Church & Hanks (1990). The resulting *I(x,y)* would be $\gg 0$. In case of no relationship between *x* and *y*, *I(x,y)* ~ 0. If the two points *x* and *y* are in complementary distribution, the probability of *P(x,y)* would be much less than 0, hence *I(x,y)* $\ll 0$ Church & Hanks (1990).

(4)

(Pointwise Mutual Information) $$I(x,y) = log_2 \frac{P(x,y)}{P(x) \times P(y)}$$

Church & Hanks (1990)

The informal way of reading Pointwise Mutual Information would be that if two points (or words) x and y have a probability of *P(x)* and *P(y)*, their mutual information *I(x,y)* would be the logarithmic quotient of the probability of both x and y by their individual probability.

In Jäger (2015) the research revolves around applying weighted string alignments in order to determine lexical and phonetic change. Jäger used string alignments to determine dissimilarities between doculects measuring the pairwise similarity between two words for the purpose of identifying whether the similarity between these words had arisen by chance Jäger (2015). He used probable cognate pairs to estimate PMI scores in order to align translation pairs with the Needleman-Wunsch algorithm with the PMI weights from the previous step. He then could deduce that all pairs above a specific threshold are cognates and hence related Jäger (2015). The figure 7 show the resulting PMI scores between ASJP sound classes.

These PMI scores were used as weights in the NW method in order to score more precise results. The PMI scores, kindly provided by Gerhard Jäger, show the likelihood of the co-occurrence of two sounds. For the algorithm, this means that in addition to the match, mismatch and gap behaviour, another measurement was included which would favour a co-occurrence of, for example, "b" and "v" rather than "a" and "r" Jäger (2015). Once these scores/distances were calculated and were corrected for word length, for each language pair, the mean distance/similarity was calculated and divided by the off-diagonal score, as was described above. These results, the NWND and

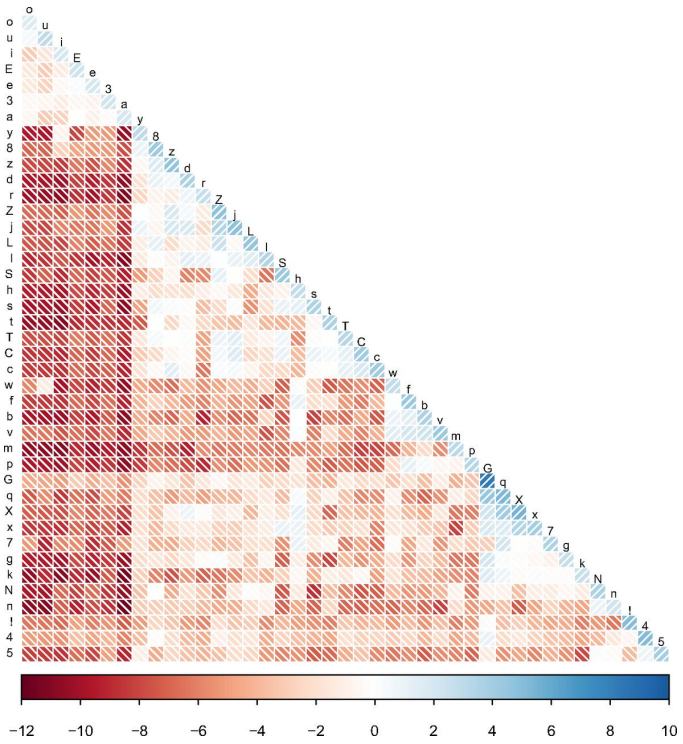


Figure 7. PMI Scores by Jäger (2015)

LDND, are the foundation for the analysis conducted by k-means clustering.

k-means clustering belongs to the group of partition clustering methods. In partition clustering methods, in opposition to hierarchical clustering, data is assigned into k number of clusters without hierarchical structure by optimisation of some criterion function. A commonly used criterion is the Euclidean Distance. *k*-means clustering takes a user-defined number of clusters k. The crucial point is to define centroids for each cluster. The function is given in *k*-means.

(5)

(Euclidean Distance) $$\delta_{r,s} = \sqrt{\sum_i (q_i - p_i)^2}^{1/2}$$

(6)

(*k*-means) $$MINIMIZE \quad J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|(x_i)^j - c(_j)\|^2$$

$\|(x_i)^j - c(_j)\|^2$ is any chosen distance measure between a data point (for instance Euclidean distance) $(x_i)^j$ and the cluster centroid Saxena et al. (2017). *k*-means calculates the object's distance to the centroids until the group object's minimum distance is calculated and can hence unarbitrarily be assigned to a cluster.

As the number of clusters *k* is not predefined, one needs to determine the optimal number of clusters. There are multiple ways of doing, this such as calculating the sum of squares at each number of possible clusters, better known as the "Elbow Method", Gap Statistics or the Silhouette Method. For the purpose of this thesis I fitted finite mixture models in order to determine the number of components of a model. This approach is also referred to as "unsupervised clustering" or "model-based clustering" Benaglia et al. (2009) and is more exact than the aforementioned methods. The implementation of the model-fitting was done with the *mixtools* package in R.

For some cases it holds that populations can be divided into subgroups. Even if the subgroups could be visually determinable (for example with a distribution graph), this approach is not very exact Benaglia et al. (2009). To specify and verify the subgroups (or components) in a given distribution, finite mixture model fitting is a helpful tool. The resulting component number can hence be used as the number of clusters in *k*-means clustering.

The chosen method for this study was to determine the number of clusters by testing models with *k*-components against *k+1*-component performing the likelihood ratio test of 100 bootstrap realisations. The bootstrapping was done in order to verify whether the data set can be held accountable. The threshold was $p < 0.05$. In this manner, the existence of 1 to 6 components was tested. In practice, this means that a model with *k* components which is more significant than another model with *k+1* components is being chosen as the most likely model. Consequently, the outcome of this fitting of mixture models, namely the number of components can be chosen as the optimal number of clusters.

The analysis was conducted in *R version 3.5.3* with the *stats* R Core Team (2019), *factoextra* Kassambara & Mundt (2017), *tidyverse* Wickham (2017) and *fpc* Hennig (2018) packages. The finite mixture-model was implemented with the *mixtool* package in *R version 3.5.3* Benaglia et al. (2009).

A further tool I used is *Gabmap* Nerbonne et al. (2011) which is a web application for visualising dialect variations on maps. Furthermore, it comes

with statistical analysis of the data and also conducts Multidimensional Scaling and clustering.

Considering the study of Søren Wichman Wichmann (2019), where two clusters were found, I nonetheless expect to find more than one objective threshold and consequently more than two clusters. The reason for this assumption is grounded in the hypothesis of dividing a language family into groups of dialects-dialects, languages-languages and "moderately distant/close pairs inbetween those stages". Furthermore, I assumed that the results with the Needleman-Wunsch algorithm will be more refined and yield better clusters and hence bear groupings of language-dialect division.

4. Results

The number of clusters was determined with the LDND and NWND data by testing models with *k*-components against *k+1*-components performing the likelihood ratio test of 100 bootstrap realisations.The threshold was $p < 0.05$. I tested 1 to 6 components and got a result of *k=3* for the Levenshtein data and *k=4* for the Needleman-Wunsch data. The resulting number of clusters is consequently 3 for the Levenshtein data and 4 for the Needleman-Wunsch data. These can be visually seen in figure 8:

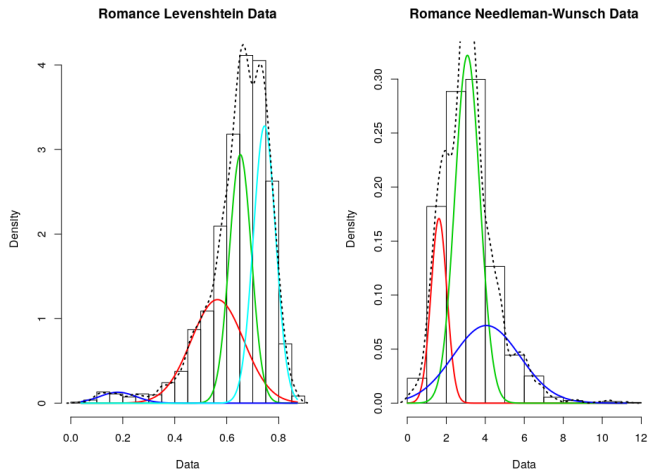


Figure 8. Romance Density Plots with 4 Density Curves in Green, Dark Blue, Light Blue and Red on the Left and 3 Density Curves in Red, Green and Dark Blue on the Right

4.1. Levenshtein Clusters

The Levenshtein clusters with their according size, centres and ranges can be seen in table 2. What can be observed is that the smallest cluster is that with the lowest distances, namely cluster 3. It almost only covers instances of true dialect-dialect pairs that can be assessed by the majority of the literature and were also manually checked.

Cluster Number	Centre	Cluster Range	Cluster Size
3	0.235	0.0394 — 0.3678	114
2	0.505	0.3765 — 0.5730	544
4	0.642	0.5740 — 0.6960	1374
1	0.75	0.6963 — 0.8713	1274

Table 2. Levenshtein Cluster Number with Centres, Size and Range of Levenshtein Value

These pairs include combinations of two varieties of Catalan, Piemontese, Venetian, Romansh, Standard Italian/other Tuscan dialects, Sicilian, Ligurian, Ladin languages, hence pairs of close varieties. This systematic almost exclusively prevails except for the pair of Asturian-Galician (ast-glg) and Standard Italian-Venice Venetian which stand out because both pairs combine two varieties which are allegedly not very similar. Apart from this "irregularity", every composed pair belongs to the same "Standard Variety". This does not mean that this automatically declares these dialects as true dialects of a given Standard variety but rather that the alleged dialects of, for instance, Standard Italian, are closest to the dialects that are spoken in their vicinity. Consider the pair of Standard Italian (ita) and Venice Venetian (ve1), Standard Italian and Grosseto Italian (itg) which are also dialects of each other by the opinion of many scholars. These results support the popular claim in the theory of Italian Dialectology that Standard Italian is rather a dialect which has been announced the official language but is by no means "superior" to other dialects of Italy. The distance of Standard Italian to other Italian varieties can be seen in 1 where an indicator is the darkness of colour. The darker the colour, the closer are the varieties. This depiction shows us some remnants of a dialect continuum which supports the hypothesis that Standard Italian is not “superior” to any other Italian variety. If the picture was predominantly dark blue, the conclusion would have been different and we could not have made out a dialect continuum as this hypothetical picture would have suggested that the Standard variety is very close to almost every

other variety. The observation is that, despite having a very big influence on other varieties, Standard Italian is also a dialect of the languages of Italy, as are all the other varieties. This influence of Standard Italian can be seen in Saladino (1990) for instance with the change of initial *f* to *h* in the local variety but the adaptation to the Standard Italian variety. The informants of the study would produce *"ficu"* (engl.: fig) instead of *"hicu"* which is the regional varietal form.

In cluster 2 we see pairs of varieties of moderately larger distances, both geographically and, naturally, also lexically. These pairings predominantly consist of two varieties of the same "Standard Variety" according to the literature. Examples for these pairings are Vercellese Piemontese-Bergamo Lombard, Primiero Venetian-Fassano Ladin, Turinese Piemontese-Rapallo Ligurian and Reggiano Emiliano-Bellunese Venetian. One can see that the cluster predominantly bears these pairs of dialects considered to belong to the same standard language, as opposed to the pairs of dialects belonging to different languages. Exceptions of this pattern are some pairings like Reggiano Emiliano-Castello de la Plana Catalan, Manises Catalan-Carpigiano Emiliano and Reggiano Emiliano-Valencia Catalan. As this pairing of an Emiliano variety with Catalan seems to be regular, one cannot assume that these are outliers where the similarity occurred by chance but rather a systematic trend which might be explainable with studying the actual word lists. As this is not part of the research question, I will not further dwell on this.

Cluster 4 does not cover those pairs that are very far away from each other but just below that threshold. The pairs in this cluster are not considered dialects in relation to each other. Examples for these pairs are Grosseto Italian-Castillan Spanish, Friulian-Turinese Piemontese and South-Eastern Sicilian-Primiero Venetian. Nevertheless, there are some pairs that can be considered dialects to each other such as Genoese Ligurian-Ferrarese Emiliano. As these pairings of ambiguous fealty are in the lower range of the cluster, it can be assumed that the algorithm did not catch all of the instances, but for the majority of these pairs, it holds that they can be considered different languages in relation to each other. This means that at the "border" of the clusters we can assume that some pairs are not entirely unambiguous, as was seen in cluster 4 where both pairs of small and big distances occurred.

The last cluster, as the range in Levenshtein scores suggests, covers those pairs in the language family that are very distant to each other. These pairs are for instance Neapolitan-Archaic Latin, Romanian-Catalan, Provencal

Occitan-Catalan and Portuguese-Plesio Lombard. It can safely be assumed that they are pairs of distinct languages.

Summarising the results from the Levenshtein analysis, three cut-off points are suggested. The first suggested threshold is 0.37 as that is the point where the cluster with the lowest distances meets the cluster with moderately low distances. With this threshold, we separate dialect-dialect pairs from dialect-language pairs. The second cut-off point is suggested at 0.58 where the cluster with the moderately low distances meets moderately higher distances. This threshold divides dialect-language pairs with moderately higher distances from other dialect-language pairs which are further apart from each other. The last cut-off point is suggested at 0.7 where the those pairs are grouped in with very high distances and hence can be seen as distant pairs in relation to each other.

4.2. Needleman-Wunsch Clusters

As visible in 3 the results from the Needleman-Wunsch clusters differ in number. This depicts a different picture of the pairing results and will hence be analysed differently. Consider the cluster pair plot on the following page. The table in 3 shows the cluster number with the according size, centres and ranges.

Cluster Number	Centre	Cluster Range	Cluster Size
1	1.7894	0.0012 — 2.5527	1093
2	3.3212	2.5643 — 4.4181	1739
3	5.5237	4.4329 — 11.2561	474

Table 3. Needleman-Wunsch Cluster Number with Centres, Size and Range of Needleman-Wunsch Value

The optimal number of clusters with the Needleman-Wunsch algorithm yields *k=3*. For the range with high similarities, it still holds that it bears the smallest cluster in size. Interestingly, the cluster with the lowest similarity is not the largest but rather the cluster that covers the "middle part" is.

Starting with the cluster bearing the highest similarities, we can see a very vast range, namely from 4.4329 to 11.2561. This is only problematic in

terms of the lowest values which are given by the pairs Sicilian-Catalan, Sicilian-Galician and Neapolitan-Portuguese for example. These pairs bear a very low similarity but are still included in the cluster with pairs of high similarity. Except for the mentioned outliers, the cluster is purely covered with dialect-dialect pairs such as varieties of Emiliano Romagnol, Venetian, Sicilian and Catalan.

Cluster 1, which bears the lowest similarities, predominantly covers language-language pairs. The highest value bears the pair of Classical Latin-Romanian for which can be said that they definitely are two distinct languages. Concluding from this value, if the the highest value is already a pair of two distinct languages, one can assume that those with lower similarity are also distant languages. Further pairs include Catalan-Romansh, Old French-Ligurian, Logudorese-Ligurian and many more. The purity of this cluster by bearing only language-language pairs can be confirmed.

The last cluster to be analysed is the cluster 2. These groupings of languages cover by far the largest cluster in the analysis. They include very distinct languages with a range from 2.5643 to 4.4181, but also pairs with higher similarity in the upper part of that range. These pairs are, amongst others, Galician-Spanish, Old Italian-Classical Latin and Venetian-Piemontese.

Summarising the results from the Needleman-Wunsch data, there are several options to analyse this. On the one hand we have predominantly pure clusters at the extreme ends. This means that the clusters with very high and very low similarities are not very noisy and their borders can be seen as natural cut-off points. On the other hand, we have a large in-between cluster which covers instances of dialect-dialect and language-dialect pairs. This could either be seen as an ambiguous result with no clear results in terms of finding a threshold or rather like a transgression between these two stages. As has been argued before, it is natural that the similarity between languages and dialects is not static but changes constantly. The importance of this in-between cluster is very high in terms of showing the trends in a language family. What this means is that language varieties are subject to diversification and convergence processes. Imagine this like a very slow wave - it moves away from you but also comes back. This metaphor should not be taken literally but rather help understand that language is not static. The system evolves continuously. Features are being adapted, sounds get borrowed, lexical items get borrowed, and sometimes they also get lost. The consequence of these processes is that varieties change constantly and either converge or diversify,

which is represented by the in-between cluster.

The sizes of this in-between cluster allow us a glimpse into the direction the different pairings "move". A direction can go either way - towards being dialect-dialect pairs or distancing themselves from each other and becoming language-language pairs. Hence it is not surprising that this cluster is the largest because, as already mentioned, language is always in movement towards either side of the extremes. In addition to that, it also happens that this movement rotates. What could happen is, for example, “true“ dialect pairs like Palermitan Sicilian and Messinese Sicilian digress from each other and become independent languages in relation to each other at some point in the future and language pairs like Palermitan Sicilian and Venice Venetian approach each other in the long run. These are not prognoses for the future; they just explain what *could* happen based on the assumption that languages change and are not static.

How are these stages manifested in the data? We see that in the lower range of the in-between cluster, the part that covering pairs with lower similarities and approaches language-language pairs, the combined varieties are pairs which are expected to be more distant such as for instance Catalan(Spain)-Piemontese(Italy) or Asturian(Portugal)-Logudorese(Italy). In the higher range of that cluster, the part that covering pairs with higher similarities and approaches dialect-dialect pairs, we see pairs like Venetian(Northeastern Italy)-Piemontese(Northwestern Italy) or Old Italian-Late Classical Latin. The mentioned pairs are not exhaustive of what is covered in this cluster but give an idea of what is being dealt with here. The pairs cover combinations of variety pairs with different degrees of similarity. This can range from distant pairs which presumably are different varieties of different “Standard Varieties” up to pairs with high similarity, presumably belonging to the same “Standard Variety”.

Drawing from all these observations, one can safely assume that the cluster with the highest similarities bears the threshold for dialect-dialect pairs. The cluster with the lowest similarity scores draws a line between language-language pairs and language-dialect groupings. As already mentioned, the cluster in the middle can be interpreted as a digression stage. Henceforth, I propose a threshold of 4.41 on a similarity scale for distinguishing dialect-dialect pairs from language-dialect pairs and a threshold of 2.54 in order to distinguish dialect-language pairs from language-language pairs.

5. Discussion

Considering the research question proposed of how to distinguish pairs of languages from pairs of dialects and the proposition in the introductory chapter to find **more than one** objective threshold and prove that the NWND results yield more precise information than LDND results we can conclude the following:

The proposition aspect can be confirmed. The results clearly showed that there is more than just one distinction between a language and a dialect. This result also seems to be adhered to if one considers Chambers et al. (1998) as the statement was that "a language is a collection of mutually intelligible dialects" which makes every variety of a language not "more of a language" or "less of a language". Language is handled as a cover term for many varieties if one decides to pick this definition. This statement argues against the notion of having a standard variety which is "superior" to their dialects. The argument by Chambers et al. (1998) is the reason why I chose to examine *pairs* of varieties and not singular varieties and their degree of "being a fully fledged language". The results seem to support that there is more than one threshold, namely (depending on the method) either 2 or 3 thresholds. In any case, the distinction that is being made is either between dialect-dialect pairs and pairs of small distances; between pairs of small distances and higher distances (in case of LDND); and between pairs of higher distances and language-language pairs. As Chambers et al. (1998) suggested that every variety of a language is equal I suggest that the in-between cluster cannot be seen as dialect-language pairs but rather as "pairs of moderately higher/lower distances in respect to the extremity points".

Regarding the second proposition I formed, this cannot be answered easily or clearly. On the one hand, LDND delivers results which already perform this partition into "pairs of moderately higher/lower distances in respect to the extremity points" with the introduction of the fourth cluster. On the other hand, the NWND results seem to already perform that partition and group the extreme points of the in-between cluster to the global extreme points (which means to the clear dialect-dialect pairs or the language-language pairs). How these results can be established and analysed can be assessed in the following discourse.

Comparing both methods there are multiple conclusions to draw. On the one hand we have the LDND results and on the other hand we have the NWND

scores. The optimal number of clusters seems to have been determined by the distribution the methods yielded, which was 4 clusters for the LDND results and 3 clusters for the NWND scores. How can these groupings be interpreted?

Considering the distributions in 9, it is unexpected that the optimal number of components yielded *k=3* for the NWND results and *k=4* for the LDND results. What could have been expected for the optimal number of components, comparing the work at hand to Wichmann (2019), is *k=2* like in Wichmann (2019). As this is not the given result in the current study, the question arises how to determine the source of this result.

Despite *seeing* 2 underlying large distributions, the data might be much more fine-grained than it appears. Considering the distribution for the LDND in 9 for instance, it is visible that the results on the left show smaller components more accentuated than the NWND results on the right. The dark blue curve in the LDND data where the pairs of languages are included with very low distances is largely absent in the NWND data. Further divisions are made, also mirrored in the cluster analyses, into pairs of very dissimilar pairs and the two groups between those points, namely pairs of moderately higher distances and moderately lower distances. For the NWND results it holds that the extreme points are also realised by the shown distributions but the transitions are more gradual. This would explain why *k=3* was chosen as the optimal number of components.

One assumption that could be made is that for each method and each distribution there is an optimal number of clusters. It could be that the LDND distance will always yield 3 clusters when given roughly the same amount of data and the same holds for NWND. This would mean that the method is the decisive driving factor for this choice of clusters. The question then still arises why in Wichmann (2019) the optimal number of clusters was only 2. This could be explained with the size of the data and the slightly different method he used. The ASJP data (40 concept word list) that Wichmann used contains fewer concepts than the data set used for the current study. This could be a legitimate reason for the difference in outcome but is rather unlikely as it has been shown that for historical inferences there is no need for word lists longer than 40 for a stable result. Furthermore, what differs greatly between both studies are the methods. Wichman used LDN distances, whereas I use LDND distances. This might have a great influence on how the distribution is affected, as the differences in the methods are as

follows: Dividing the mean value of a language pair of either Levenshtein Distances or Needleman-Wunsch scores by the mean of the off-diagonal score of that specific language pair accounts for the phoneme inventory in those languages. This modulation ensures that languages with a very high overlap in their phoneme inventories do not get a low distance/high similarity score in cases where it is not appropriate. These cases of improper accounting of a low distance/high similarity score could happen when the alignment does not account for that. Following this logic, the safer way of looking for distances or similarities between languages is by not only correcting for word length but also for their phonemic inventories Jäger (2014).

The assumption I make is that the combination of the method and the data size plays a major role in how the outcome is going to look like. It cannot be ruled out that the data set size is a crucial driving factor for the outcome of the studies. As the only overlaps with this study and Wichmann (2019) are the partial data set and the method in terms of its fundamental operating principle, there cannot be said anything further in terms of the interpretation of the different results. One could nonetheless test either the LDND method with the exact data set Wichmann (2019) used, or apply the LDN method to the data set used for this study.

Another question that could be asked is why the methods used in this study deliver different clustering results than those of Wichmann (2019). Reconsider the plot in 9.

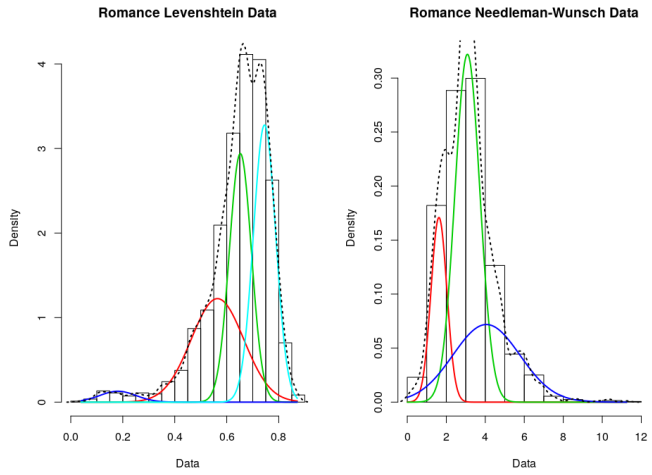


Figure 9. Romance Density Plots

One possible reason in order to explain why the results from the cluster analysis by Wichman and this study differ might be due to the already mentioned subtleties that are caught by dividing the mean score by the off-diagonal score. If one considers the histograms in 9, it is visible that the main difference between the distributions in the current study and the distribution for Iranian in Wichman's study (which looks like a bimodal distribution) are the nuances shown in the different scores. This means that rather than having extreme points on the LDND and NWND scale, this study bears many gradual scenarios between said extreme points. As already mentioned before, a language family naturally does not yield only dialect-dialect and language-language pairs but also "pairs of moderately higher/lower distances in respect to the extremity points". These were accounted for with NWND/LDND results. This leads me to the assumption that the method used in Wichmann (2019) is not entirely wrong and also gives the scholar a first idea about the relationships between the different doculects in a language family, but is not as well suited for an analysis of this kind as the NWND/LDND method.

Another remarkable observation in the histograms is the distribution of the data in comparison. The distribution for the LDND data follows a bimodal distribution. This could lead to the assumption that LDND draws a sharper cut-off by nature. The presumption that the language families have different distributions when looking at pairs and their distances is hence more likely. This brings me to the next suggestion, which is the consultation of several data sets with the affirmed NWND method in order to draw a clearer picture of the language family individually.

Concluding the results given in the previous section and also considering the thoughts and interpretations of these results, the question arises which method to opt for and what the suggestion for future research might be. While the LDND method gave unexpected results, it provided a partition in the in-between cluster between pairs of moderately high and moderately low/high distances/similarities. The NWND data, however, delivered a tripartite division which shows larger extreme groups and smaller in-between groups. The division into three groups is reasonable if one consults an analysis of this kind for the partition into discrete groups, whereas the division into four clusters is reasonable for the purpose of dividing the in-between cluster.

The final assessment is to opt for the NWND method with a weighted scorer system in order to obtain to obtain results which show discrete groupings. For a more detailed analysis, the LDND results serve better and provide a

clearer explanation of the in-between cluster. It follows that for the analysis of the extreme points, NWND delivers a clearer results and for the analysis of the in-between cluster, LDND is the method to opt for.

6. Conclusion

Summarising the findings, I can conclude that between the two methods of Levenshtein Distance Normalised and Divided (LDND) and the Needleman-Wunsch algorithm Normalized and Divided (NWND), the results provided by NWND yielded better results in terms of analysing the extreme ends of the clusters. Albeit not being optimal in determining the "pairs of moderately higher/lower distances in respect to the extremity points", NWND gave us more expressive results and in order to answer the research question posed in this study. The suggested threshold by the NWND method are 4.49 for distinguishing dialect-dialect pairs from "pairs of moderately higher/lower distances in respect to the extremity points" and a threshold of 2.54 in order to distinguish "pairs of moderately higher/lower distances in respect to the extremity points" from language-language pairs. For the LDND method the cut off-points are 0.37 to distinguish dialect-dialect pairs from "pairs of moderately higher/lower distances in respect to the extremity points", 0.58 to distinguish close "pairs of moderately higher/lower distances in respect to the extremity points" from distant "pairs of moderately higher/lower distances in respect to the extremity points" and 0.7 to distinguish distant "pairs of moderately higher/lower distances in respect to the extremity points" from language-language pairs.

This approach of distinguishing between languages and dialects with distances and similarities is a legitimate one but by far not the most expressive one. This study can be seen as a pilot study for further investigations in the field.

Another possible step for future research could be to change the nature of the entries. As ASJP and The Global Lexicostatistical Database worked with Swadesh lists, it might probably be the case that other concepts yield a better understanding of the languages development, as for instance in the Romance language families, many concepts for the basic vocabulary overlap in most of the languages. As the words for the most basic vocabulary usually do not undergo major shifts, they are well preserved which gives a good indication of genealogical relations but for the fine nuances in determining dialects, the

suggested approach could yield clearer outcomes.

Another important aspect, which was entirely disregarded in this study, is the analysis of syntactic properties in languages. This might be a great indicator of how close languages and dialects are and could be consulted in addition to a phonemic and lexical analysis.

APPENDIX

Varieties and their Individual Codes

Variety	Code	Variety	Code
qbb	Archaic Latin	ve2	Primiero Venetian
lat	Late Classical Latin	ve3	Bellunese Venetian
ruq	Megleno Romanian	ito	Old Italian
ruo	Istro Romanian	ita	Standard Italian
rup	Aromanian	itg	Grosseto Italian
ron	Romanian	itf	Foligno Italian
dlm	Dalmatian	nea	Neapolitan
fur	Friulian	src	Logudorese
ll1	Gardenese Ladin	sro	Campidanese
ll2	Fassano Ladin	sc1	Palermitan Sicilian
ro1	Rumantsch Grischun	sc2	Messinese Sicilian
ro2	Sursilvan Romansh	sc3	Catanian Sicilian
ro3	Surmiran Romansh	sc4	South-Eastern Sicilian
ro4	Vallader Romansh	ca1	Central Catalan
pm1	Lanzo Torinese Piemontese	ca2	North-Western Catalan
pm2	Barbania Piemontese	ca3	Minorcan Catalan
pm3	Carmagnola Piemontese	ca4	Castello de la Plana Catalan
pm4	Turinese Piemontese	ca5	Valencia Catalan
pm5	Vercellese Piemontese	ca6	Manises Catalan
lm1	Bergamo Lombard	spa	Castilian Spanish
lm2	Plesio Lombard	ast	Asturian
eg1	Ravennate Romagnol	por	Standard Portuguese
eg2	Ferrarese Emiliano	glg	Galician
eg3	Carpigiano Emiliano	pr1	Provencal Occitan
eg4	Reggiano Emiliano	pr2	Savoyard Franco-Provencal
li1	Rapallo Ligurian	fro	Old French
li2	Genoese Ligurian	fra	Standard French
li3	Stella Ligurian	pcd	Picard
ve1	Venice Venetian	win	Walloon

REFERENCES

Benaglia, Tatiana, Didier Chauveau, David R. Hunter & Derek Young. 2009. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software* 32(6). 1–29.

Brown, Cecil H, Eric W Holman, Søren Wichmann & Viveka Velupillai. 2008. Automated classification of the world's languages: a description of the method and preliminary results. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung* 61(4). 285–308.

Chambers, J.K., P. Trudgill & S.R. Anderson. 1998. *Dialectology*, Cambridge Textbooks in Linguistics. Cambridge University Press.

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16(1). 22–29.

Clivio, Gianrenzo P, Marcel Danesi & Sara Maida-Nicol. 2011. *An introduction to italian dialectology*. Lincom Europa.

Durbin, R., S.R. Eddy, A. Krogh & G. Mitchison. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press.

Goebl, Hans. 1982. *Dialektometrie*, vol. 157. Verlag der Österreichischen Akademie der Wissenschaften.

Gooskens, Charlotte. 2006. Linguistic and extra-linguistic predictors of inter-scandinavian intelligibility. *Linguistics in the Netherlands* 23(1). 101–113.

Gusfield, D. 1997. *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge University Press.

Heeringa, Wilbert Jan. 2004. *Measuring dialect pronunciation differences using levenshtein distance*. Ph.D. thesis, Citeseer.

Hennig, Christian. 2018. *fpc: Flexible procedures for clustering*. R package version 2.1-11.1.

Jäger, Gerhard. 2014. Phylogenetic inference from word lists using weighted alignment with empirically determined weights. In *Quantifying language dynamics*, 155–204. Brill.

Jäger, Gerhard. 2015. Support for linguistic macrofamilies from weighted sequence alignment. *Proceedings of the National Academy of Sciences* 112(41). 12752–12757.

Kassambara, Alboukadel & Fabian Mundt. 2017. *factoextra: Extract and visualize the results of multivariate data analyses*. R package version 1.0.5.

Kessler, Brett. 1995. Computational dialectology in irish gaelic. In *Proceed-*

ings of the seventh conference on european chapter of the association for computational linguistics, EACL '95, 60–66. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

King, Robert D. 2001. The poisonous potency of script: Hindi and urdu. *International journal of the sociology of language* 2001(150). 43–59.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, vol. 10, 8, 707–710.

List, Johann-Mattis. 2013. *Sequence comparison in historical linguistics*. Ph.D. thesis, Heinrich-Heine-Universität Düsseldorf.

List, Johann-Mattis, Simon Greenhill & Robert Forkel. 2017. Lingpy. a python library for quantitative tasks in historical linguistics.

Maiden, M. & M.M. Parry. 1997. *The dialects of italy*, Romance Linguistics. Routledge.

Needleman, Saul B & Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48(3). 443–453.

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg & Therese Leinonen. 2011. Gabmap-a web application for dialectology. *Dialectologia: revista electrònica* 65–89.

R Core Team. 2019. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saladino, Rosa. 1990. Language shift in standard italian and dialect: A case study. *Language Variation and Change* 2(1). 57–70.

Saxena, Amit, Mukesh Prasad, Akshansh Gupta, Neha Bharill, Om Prakash Patel, Aruna Tiwari, Meng Joo Er, Weiping Ding & Chin-Teng Lin. 2017. A review of clustering techniques and developments. *Neurocomputing* 267. 664–681.

Starostin, George S. 2011. 2015. the global lexicostatistical database. moscow/santa fe: Center for comparative studies at the russian state university for the humanities; santa fe institute.

Swadesh, Morris. 1971. *The origin and diversification of language*. Transaction Publishers.

Wichmann, Søren. 2019. How to distingusih languages and dialects. *Computational Linguistics* 45.4. 823–831.

Wickham, Hadley. 2017. *tidyverse: Easily install and load the 'tidyverse'*. R package version 1.2.1.

Wieling, Martijn & John Nerbonne. 2015. Advances in dialectometry .